DOCUMENT RESUME

ED 295 965 TM 011 750

AUTHOR Drahozal, Edward C.
TITLE Pupil Norms or Norms for School Averages: Vintage Means of Test Interpretation or a Can of Worms?
PUB DATE Apr 88
NOTE 31p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; College Entrance Examinations; *Educational Planning; Elementary Education; Grade 2; Grade 6; Grade 8; *National Norms; *Standardized Tests; *Test Interpretation; Test Use
IDENTIFIERS American College Testing Program; *Iowa Tests of Basic Skills; *Stanford Achievement Tests

ABSTRACT

The purposes of this study were to compare student, school, and group average distributions of two standardized tests, the Iowa Tests of Basic Skills (ITBS) and Stanford 7 Plus Achievement Test, and examine the American College Testing Programs college entrance examinations (Grades 11-12). In addition to the student-school norm comparisons, ratios of school-student norms were examined. ITBS and Standard 7 data on second, sixth, and eighth graders were analyzed. Results indicate that: (1) single-point indicators of achievement have little, if any, value to instructional planners; (2) since the percentile ranking of a student is reported to the student and his or her parents, group data should be reported via group norms rather than student norms; (3) single-point indicators are often misrepresented and misused, particularly by reporters; and (4) "correlates" of achievement are rather imperfect, but probably vary more from state to state than within a state. Tests developed at the local and state levels do not adhere to national standards concerning supporting technical data; these standards are met by most commercially available tests. Seven tables are presented, and a speech by E. F. Lindquist (1966) is appended. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED295965

Pupil Norms or Norms for School Averages: Vintage Means of Test Interpretation or a Can of Worms?

Paper Presented
at the
American Educational Research Association
New Orleans
April 1988

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

EDWARD C. DRAHOZAL

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

by

Edward C. Drahozal
The Riverside Publishing Company
8420 Bryn Mawr Avenue
Chicago, IL 60631
800/323-9540

Descriptors

Test Norms
Test Interpretation/Use
Test Reporting

TM 011 750

ABSTRACT

Pupil Norms or Norms for School Averages:
Vintage Means of Test Interpretation or a Can of Worms?

Edward C. Drahozal

Pupil norms are often used to report percentile ranks for group averages. A recent Friends For Education report stated that all states were above average, but in reality referenced state averaged scores to the pupil median. In the national average the report ignored the differences in the percent of students tested in the various states. Critics of schools often cite socio-demographic characteristics which purportedly predict school performance when criticizing the educational program in a school or district.

This paper summarizes the following:

a. The relationship between pupil norms and norms for groups for the Iowa Tests of Basic Skills, Form G and the Stanford 7 Plus Achievement Test.

b. The relationship between national means and pupil medians for these two tests and the California Achievement Tests, Form E.

c. The amount of change in the pupil percentile ranks of mean scores for these three tests when students are systematically excluded from the groups for which the mean scores are reported.

d. The relationship between pupil and school norms for ACT and various school and demographic characteristics for Chicago suburban schools.

e. A method for adjusting pupil percentile ranks for school averages when the national pupil average and pupil median differ.

Descriptors

Test Norms
Test Interpretation/Use
Test Reporting

Pupil Norms or Norms for School Averages:
Vintage Means of Test Interpretation or a Can of Worms?

INTRODUCTION

In the 1985 Standards for Educational and Psychological Testing, which supersede the 1974 standards. Standard 4.5 reads, "When it is expected that a test will be used to make norm-referenced assessments of groups rather than individuals, normative data based on appropriate groups statistics should be provided. (Secondary)." A similar standard in the 1974 Test Standards was labeled "Essential."

An "Essential" label in the 1974 standards indicated that group norms should be used whenever the normative data available applied to the group to be described. The 1985 label, "Secondary", means to quote the 1985 standards, "desirable goals but are likely to be beyond reasonable expectation in many situations." Thus, the standard, previously labeled "Essential", which in the 1985 standards terminology would be called "Primary", is likely to be ignored. It appears that the label "Conditional" should have been assigned this standard, but the "Conditional" standard, described in the early drafts of the Standards as one which, "should be considered primary for some situations and secondary for others," was dropped from the final standards.

A recent Friends for Education (FFE) report listed data for several different kinds of scores. The FFE report claimed that all states are above average, that perhaps as many as 70% of the pupils in the United States are being told they perform above average, and that perhaps 90% of the districts in the United States are reporting that they score above average. Means of various scores are used interchangeably in the FFE report without regard to the relationship among pupil means, pupil medians, and the distributions of group averages.

HISTORICAL PERSPECTIVE

In a 1948 speech to the measurement community, E. F. Lindquist suggested that such norms are reasonable and should be provided in order to compare group achievement. He noted that the relationship between pupil variability and school variability varied considerably by test and grade. He also noted that this lack of constant variability could result in some rather serious misinterpretations of group data; that is, "the school is relatively low in math compared to language," when in fact, this is not the case. His vintage speech was largely ignored, even after it was published by Anastasi in 1966. In a recent discussion of norms at a state research meeting, one of the speakers referred to school average norms as, "a can of worms."

CURRENT PRACTICES

Current practices of reporting school test data vary by school district, depending on (1) the averages obtained and (2) the availability of published norms. No publishers provide norms for system averages. Since several publishers do not provide school average norms, pupil norms are used by many district administrators to report school data. When school average norms are available, above average schools tend to use them because school average percentile rank is greater (better, higher, etc.) than the pupil percentile rank; below average schools tend to use pupil percentile ranks because they are greater (higher) than school average percentile ranks.

Some districts and states use school average percentile ranks for reporting data for special groups such as Chapter 1, state averages, and population subgroups. Other districts and states report pupil percentile ranks. Some districts use medians and report pupil percentile ranks for school and district median scores.

PURPOSES OF THIS PAPER

The purposes of this paper are to compare pupil, school (building), and group average distributions of two standardized tests, the Iowa Tests of Basic Skills (ITBS) and Stanford 7 Plus (Stanford), that were obtained during two different time periods, and look at a test which recently began providing school norms, the American College Testing (ACT) Programs college entrance test. In addition to the pupil-school norm comparisons, ratios of school/pupil norms will be examined.

This paper will also examine differences between the means and medians of pupil scores and the effects on mean scores when groups of students are systematically or unsystematically excluded from data reported by schools. Data are provided for ITBS, Stanford 7 Plus, and the California Achievement Tests, Form E (CAT). ACT group norms will also be related to demographic characteristics in a small separate study.

DATA PROVIDED

The ITBS has pupil norms and school norms. The pupil norms are derived from a distribution of pupil scores from the national standardization; the school norms (norms for school averages) are derived from a distribution of school averages from the national standardization. The Stanford has two sets of group norms. One is for groups of 50 or fewer pupils; the second is for groups of more than 50 pupils. Neither publisher provides norms for district averages, probably for the reasons noted in the ITBS Manual for School Administrators (MSA). (See page 34 of ITBS MSA Forms G and H, 9-50280.)

The ACT group norms are based on schools with 40 or more students. Hence, they are more or less comparable to the size of the groups in the Stanford large group norms.

The school average norms for the ITBS and the two sets of group norms for the Stanford have two characteristics in common. They are less variable than pupil norms and the median of the group averages does not correspond to the pupil median in most grades on most tests.

The procedure used to obtain the comparisons between the selected pupil percentile ranks and group percentile ranks was simple but time-consuming. For the selected PR's, appropriate tables were used to obtain the SAT scaled score (SS), ITBS GE, or 1987 ACT standard score (SS). For the Grade 2 Fall SAT Reading Comprehension test you end up with a table like this.

Small Group PR	SS	Pupil PR	SS	Large Group PR
99++	667.	99	667.	99++
99	610.	88		
		80	590.	99
	615.	90	615.	
90	585.	77		
		70	572.5	90
	580.5	75	580.5	
75	568.	68		
		62	559.5	75
	541.5	50	541.5	
50	549.5	55		
		53	546.	50
	502.5	25	502.5	
25	531.	43		
		44	532.	25
	467.	10	467.	
10	514.	32		
		35	519.	10
		24	501.	1
1	489.	18		
1--	415.	1	415.	1--

The SS's (or ITBS GE's) were read to the nearest tenth, using linear interpolation methods, and all PR's were read to the nearest whole PR. These tables can be reduced to the tables shown in Table 1.

From this table and the data in Table 1, it is clear the median of group averages seldom coincides with the pupil median. The Stanford large and small group medians differ quite markedly, as one would expect.

TABLE 1: Pupil and School Percentile Ranks with Grade Equivalent Scores, Iowa Tests of Basic Skills

Reading Comprehension

Grade 2 Spring			Grade 6 Fall			Grade 8 Spring		
GE	Pupil	School	GE	Pupil	School	GE	Pupil	School
57	99		101	99		129	99	
41	86	99	82	88	99	114	91	99
44	90		84	90		112	90	
36	75	90	73.5	74	90	100	72	90
36	75		74	75		101.5	75	
33	67	75	68.8	64	75	94.8	62	75
27.8	50		62	50		88	50	
29.4	55	50	63.4	53	50	88.8	51	50
20.3	25		50	25		72	25	
25.7	42	25	57.7	41	25	80.7	38	25
15	10		40.5	10		59	10	
22.3	33	10	52.5	30	10	74	29	10
K8	1		27	1		40	1	
18	17	1	45	16	1	65	16	1

Spelling

Grade 2 Spring			Grade 6 Fall			Grade 8 Spring		
GE	Pupil	School	GE	Pupil	School	GE	Pupil	School
56	99		101	99		129	99	
41	84	99	80	82	99	109	84	99
45	90		86	90		114	90	
36	73	90	73	71	90	100.5	71.5	90
37	75		75	75		102.5	75	
32.9	65	75	68	63	75	93.7	61	75
28	50		62	50		87	50	
29.4	55	50	62.5	51	50	87.6	51	50
21.3	25		49.5	25		69	25	
26	43	25	57.5	41	25	81.8	42	25
16	10		39	10		53	10	
23.5	33	10	53	32	10	76.5	34.5	10
K8	1		22	1		33	1	
18	14	1	46	19	1	67	23	1

Usage and Expression

Grade 2 Spring			Grade 6 Fall			Grade 8 Spring		
GE	Pupil	School	GE	Pupil	School	GE	Pupil	School
62	99		103	99		131	99	
47	84	99	86	88	99	115	90	99
51.5	90		88	90		115	90	
39.5	71	90	76.5	75	90	105	76	90
41.5	75		76.5	75		104.5	75	
35	62	75	73.7	70	75	97.7	65	75
29.5	50		62	50		88	50	
30.2	51	50	63.2	52	50	89.3	51	50
19	25		47	25		68	25	
25	39	25	56.3	40	25	80.7	40	25
12	10		35	10		54	10	
20.5	28	10	50	30	10	72	30	10
6	1		23	1		35	1	
15	16	1	42	18	1	60	16	1

Language Total

Grade 2 Spring			Grade 6 Fall			Grade 8 Spring		
GE	Pupil	School	GE	Pupil	School	GE	Pupil	School
56	99		98	99		126	99	
43	78	99	83	88	99	113	90	99
46	90		84	90		113	90	
37	66	90	74	75	90	102	74	90
37.5	75		74	75		102.5	75	
33.8	59	75	68.5	65	75	95	63	75
29	50		61.5	50		87	50	
30.1	51	50	63.2	53	50	89	53	50
22	25		49.5	25		70	25	
26.3	42	25	57.8	41	25	82	42	25
17.3	10		40	10		57	10	
23.2	35	10	53	32	10	76	33	10
12	1		30	1		44	1	
19	25	1	46	19	1	66	19	1

TABLE 1: Pupil and School Percentile Ranks with Grade Equivalent Scores, Iowa Tests of Basic Skills

Math Concepts

Grade 2 Spring			Grade 6 Fall			Grade 8 Spring		
GE	Pupil	School	GE	Pupil	School	GE	Pupil	School
53	99		98	99		129	99	
37	80	99	79	87	99	108	85	99
40.5	90		81	90		112	90	
32.8	65	90	71.7	74	90	100.5	72	90
35.3	75		72	75		102	75	
30.8	58	75	67.6	65	75	94.8	61	75
29	50		61.7	50		89	50	
28.8	49	50	62.8	52	50	88.8	49	50
23	25		52.7	25		75.5	25	
26.4	38	25	58	39	25	82	37	25
18.3	10		45	10		64	10	
24.4	31	10	54	28	10	75.5	25	10
13	1		33	1		48	1	
21	17	1	48	15	1	69	15	1

Math Computation

Grade 2 Spring			Grade 6 Fall			Grade 8 Spring		
GE	Pupil	School	GE	Pupil	School	GE	Pupil	School
48	99		89	99	117	99		
37	85	99	73	87	99	101	89	99
39	90		74.5	90		101.5	90	
32.8	70	90	68	72	90	94	71	90
33.8	75		69	75		95.3	75	
30.8	60	75	64.8	60	75	90.8	61	75
29	50		61.7	50		87	50	
28.8	49	50	61.8	50	50	87	50	50
23.8	25		54.7	25		76	25	
26.4	37	25	58.4	38	25	81.6	37	25
19.5	10		47	10		65	10	
24.4	28	10	55	26	10	77	27	10
13	1		34	1		44	1	
21	14	1	49	12	1	68	13	1

Social Studies

Grade 2 Spring			Grade 6 Fall			Grade 8 Spring		
GE	Pupil	School	GE	Pupil	School	GE	Pupil	School
72	99		110	99		136	99	
52	86	99	85	82	99	109	80	99
56	90		94	90	120	90		
41.5	75.5	90	75	71	90	100	70	90
41	75		78	75		104	75	
34.3	64	75	68.3	61	75	93.3	59	75
28	50		62	50		87	50	
28.7	52	50	61.5	49	50	86.3	49	50
19.3	25		45.5	25		66	25	
24.6	40	25	53	35	25	75	35	25
11	10		33	10		50	10	
21	30	10	45	24	10	64	23	10
K3	1		20	1		33	1	
16	17	1	36	14	1	53	12	1

Science

Grade 2 Spring			Grade 6 Fall			Grade 8 Spring		
GE	Pupil	School	GE	Pupil	School	GE	Pupil	School
68	99		111	99		138	99	
47	84	99	87	81	99	114	80	99
54	90		98	90		125	90	
40	74	90	79	72	90	105.5	70.5	90
40.5	75		82	75		110	75	
33.8	64	75	70	61	75	96.5	60.5	75
28	50		62	50		87	50	
29.2	53	50	62	50	50	86.5	49.5	50
17	25		44	25		64	25	
25.2	43	25	54.3	39	25	76.5	38.5	25
8.5	10		32	10		48	10	
20.5	32	10	48	30	10	68	29	10
10	1		19	1		32	1	
13	17	1	38	17	1	56	17	1

TABLE 1 - Stanford 7 Plus Achievement Test

Pupil, Large Group, and Small Group Percentile Ranks With Scaled Scores

Reading Comprehension

Grade 2 Spring				Grade 6 Fall				Grade 8 Spring			
SS	Pupil	Large	Small	SS	Pupil	Large	Small	SS	Pupil	Large	Small
683	99			759	99			781	99		
634	88		99	708	87		99	734	88		99
617	79	99		688	76	99		721	81	99	
639.5	90			715	90			739	90		
612	76		90	687	75		90	709	74		90
603	70	90		674	66	90		705	71	90	
610.5	75			686.5	75			711	75		
597	66		75	673	65		75	694	62		75
592.5	62+	75		664.5	58+	75		693.5	61+	75	
577	50			654.5	50			679.5	50		
581	53		50	658	53		50	677	48		50
580	52	50		652	48	50		680	50	50	
544	25			622	25			647.5	25		
565	40		25	643	41		25	659	33		25
567	42	25		639.5	37+	25		666	39	25	
515	10			593.5	10			619	10		
551	30		10	629	30		10	645	23		10
557	34	10		630	31	10		654	29	10	
471	1			550	1			577	1		
528	16		1	608	17		1	619	10		1
542	24	1		616	21	1		638	19	1	

Spelling

Grade 2 Spring				Grade 6 Fall				Grade 8 Spring			
SS	Pupil	Large	Small	SS	Pupil	Large	Small	SS	Pupil	Large	Small
679	99			776	99			808	99		
629	88		99	720	86		99	757	87		99
612	79	99		698	74	99		730	71	99	
635	90			730	90			764	90		
605	75		90	698.5	74.5		90	734	74		90
595	68	90		684	65	90		720	64	90	
605.5	75			699	75			735.5	75		
590	64		75	682	63		75	719	63		75
585.5	60+	75		674.5	57.5	75		712.3	58	75	
572.3	50			664.5	50			702.5	50		
574	51		50	667	52		50	702.5	50		50
573	50	50		663.7	49.7	50		704	51	50	
538.5	25			630.5	25			670	25		
557	38		25	652	40		25	686	37		25
560.5	40+	25		653	41	25		695.7	44	25	
509	10			599	10			641	10		
543	28		10	638	30		10	671	26		10
551	33	10		643	33	10		688	38	10	
465	1			553	1			597	1		
519	14		1	615	17		1	648	13		1
534	22	1		629	24		1	677	30	1	

TABLE 1 - Stanford 7 Plus Achievement Test

Pupil, Large Group, and Small Group Percentile Ranks With Scaled Scores

Language

Grade 2 Spring				Grade 6 Fall				Grade 8 Spring			
SS	Pupil	Large	Small	SS	Pupil	Large	Small	SS	Pupil	Large	Small
No Test				740	99			778	99		
				702	87		99	729	88		99
				687	77	99		719	78	99	
				707.5	90			740.5	96		
				684	75		90	710	71		90
				675	66	90		706.5	68+	90	
				684.5	75			715	75		
				673	64		75	697.5	60+		75
				666	57	75		697	60	75	
				659	50			686	50		
				661	52		50	683	47		50
				657	47	50		687	51	50	
				634.5	25			657	25		
				649	39		25	669	34		25
				648	38	25		677.5	41+	25	
				611.5	10			632.5	10		
				638	28		10	656	24		10
				638.5	28+	10		668	34	10	
				578	1			594	1		
				620	14		1	637	12		1
				626	18	1		654	23	1	

Language Total

Grade 2 Spring				Grade 6 Fall				Grade 8 Spring			
SS	Pupil	Large	Small	SS	Pupil	Large	Small	SS	Pupil	Large	Small
No Total				746	99			782	99		
				707	88		99	738	86		99
				687	75	99		719	74	99	
				710.5	90			745.5	90		
				689	76		90	718	73		90
				676	66	90		704	61	90	
				687	75			721	75		
				677	67		75	706	63		75
				667	57	75		695	52	75	
				660	50			693	50		
				664	54		50	692	49		50
				658	48	50		684	41	50	
				633	25			665.5	25		
				651.5	41+		25	678.5	35		25
				649	39	25		672.5	30.5	25	
				610.5	10			641.5	10		
				639	30		10	666	25		10
				639.5	30+	10		663	23	10	
				575	1			604	1		
				621	16		1	647	13		1
				629	22	1		649	14	1	

TABLE 1 - Stanford 7 Plus Achievement Test

Pupil, Large Group, and Small Group Percentile Ranks With Scaled Scores

Concepts of Number

Grade 2 Spring				Grade 6 Fall				Grade 8 Spring			
SS	Pupil	Large	Small	SS	Pupil	Large	Small	SS	Pupil	Large	Small
661	99			737	99			812	99		
616	88		99	695	87		99	758	86		99
605	82	99		680	77	99		747	80	99	
620.5	90			701.5	90			768	90		
595	75		90	677	75		90	733.5	71+		90
588	70	90		666	65	90		730	69	90	
594.5	75			677	75			738.5	75		
582	65		75	664.5	63+		75	719	61		75
578	62	75		656.5	55+	75		719	61	75	
564	50			651	50			705.5	50		
567	52		50	651.5	50+		50	701	46		50
565.5	51+	50		646	45	50		706.5	50+	50	
534	25			625	25			672.5	25		
552.5	39+		25	638.5	37+		25	684	33		25
553	40	25		635.5	34+	25		694	41	25	
507	10			600.5	10			643	10		
539	29		10	626	26		10	669	23		10
542	31	10		626	26	10		683	32	10	
467	1			565	1			599	1		
518	15		1	608	14		1	646	11		1
525	19	1		612	16	1		666	21	1	

Mathematics Computation

Grade 2 Spring				Grade 6 Fall				Grade 8 Spring			
SS	Pupil	Large	Small	SS	Pupil	Large	Small	SS	Pupil	Large	Small
677	99			744	99			812	99		
627	87		99	696	85		99	755	84		99
613	80	99		686	79	99		748	80	99	
634	90			706	90			768	90		
605.5	75		90	677.5	72+		90	734	71		90
597	69	90		670	67	90		731	69	90	
605	75			681	75			740	75		
591	69		75	665.5	62+		75	719	60		75
586.5	61	75		661	59	75		720.5	61+	75	
572.5	50			652	50			706.5	50		
574.5	51		50	652	50		50	702.5	46+		50
574	51	50		648.5	46+	50		708	51	50	
540	25			624	25			673	25		
558.5	38+		25	638.5	36+		25	686	34		25
561.5	41	25		636.5	34+	25		695.5	41+	25	
511	10			599	10			643	10		
543.5	27+		10	626	26		10	670.5	23+		10
551	33	10		627	27	10		684	33	10	
468	1			561	1			599	1		
522	15		1	607	14		1	649	12		1
534	21	1		611	16	1		667	21	1	

TABLE 1 – Stanford 7 Plus Achievement Test

Pupil, Large Group, and Small Group Percentile Ranks With Scaled Scores

Environment
Grade 2 Spring

SS	Pupil	Large	Small
658	99		
630	90		99
619	82	99	
630	90		
612	75		90
607	70	90	
612	75		
602	65		75
598.5	61.5	75	
587	50		
590	53		50
588.7	52	50	
562	25		
578	41		25
579	42	25	
544	10		
568	31		10
570	33	10	
515	1		
550	14		1
558	21		1

Social Science
Grade 6 Fall

SS	Pupil	Large	Small
727	99		
688	88		99
669	76	99	
691.5	90		
669.5	76+		90
656.5	65+	90	
668	75		
657.5	66+		75
647.5	56+	75	
641	50		
645	54		50
638.3	47+	50	
614	25		
632	41		25
629.5	38+	25	
590.5	10		
620	30		10
620.5	30	10	
555	1		
602	16		1
608	20	1	

Social Science
Grade 8 Spring

SS	Pupil	Large	Small
778	99		
729	85		99
725	83	99	
739	90		
709	72		90
710	73	90	
713	75		
695	61		75
699	64	75	
682	50		
680	48		50
684.5	52+	50	
652	25		
665.5	35+		25
670	39	25	
626	10		
652	25		10
659	30	10	
586	1		
632	13		1
644	19	1	

Environment
Grade 2 Spring

SS	Pupil	Large	Small
See Above			

Science
Grade 6 Fall

SS	Pupil	Large	Small
729	99		
690	87		99
679	80	99	
695.5	90		
672	74		90
665	67	90	
673	75		
661	63		75
654.5	56+	75	
648	50		
649	51		50
645.3	47+	50	
622	25		
636.5	38+		25
635.5	37+	25	
600.5	10		
625	27		10
626	28	10	
568	1		
607	14		1
612	18	1	

Science
Grade 8 Spring

SS	Pupil	Large	Small
763	99		
719	86		99
713	82	99	
727.5	90		
701	73		90
699	71	90	
703.5	75		
688.5	61+		75
689.5	62+	75	
677	50		
675.5	48+		50
678.7	51+	50	
650.5	25		
662	35		25
668	41	25	
626.5	10		
650	25		10
658	31	10	
591	1		
632	13		1
644	20	1	

Table 1 (Cont'd)

American College Testing Program
Composite Score, 1987 Seniors

SS	Pupil PR	School PR
31	99	
22.8	73	99
26.5	90	
21.4	64	90
23.1	75	
20.4	60	75
18.9	50	
19.2	52	50
14.4	25	
17.8	44	25
10.8	10	
16.4	35	10
6	1	
9	5	1

In order to more easily view the comparisons between pupil and group norms, Table 2 was extracted from Table 1. The pupil PR is listed in the body of Table 2. the corresponding Stanford and ITBS group PR's are listed at the top of the table for each type of norm. The Stanford large group norms are listed first. Groups of 50 or more students are fairly common for many school buildings and even most small school districts.

What do the data in Table 2 indicate? First, there are some fairly substantial differences in the range of pupil PR's across tests and grades for either of the three norms. The Grade 8 pupil PR range is most often greater than the pupil PR range for Grade 2 and Grade 6. This may come as a surprise to some people who assume larger middle schools and junior high schools would result in a smaller pupil PR range for the group averages in Grade 8 than lower grades. Also, the Grade 8 distributions are no more leptokurtic than the distributions in Grade 2 and Grade 6, as indicated by the ranges of pupil PR's between the group PR of 25 and 75 in relation to the total range of scores.

For the ITBS, the range of pupil PR's is somewhat greater than for the Stanford large group norms. This is not unexpected in that all size schools are included in these norms. The range is typically about 70 to 75 pupil PR's, which is similar to the range of pupil PR's for the Stanford small group norms.

Finally, the table clearly indicates that, if you use pupil PR's to report large group averages, a PR above 80 or a PR below 20 will seldom be reported!

The ACT distributions are quite different from the ITBS or Stanford distributions. The group PR of 99 corresponds to a pupil PR of 73, which is comparable to Stanford's lowest pupil PR in Grade 8. However, the ACT scores are based on groups of 40 or more, so a pupil PR of above 80 might have been expected. The differences at the bottom are even more dramatic. A group PR of 1 corresponds to a pupil PR of 5, as compared to pupil PR's of 13 to 21 for the ITBS and Stanford Grade 8. The ACT distribution is far more leptokurtic than any of the ITBS or Stanford distributions. [The range of scores between the 25th and 75th percentile on group norms (17.8 to 20.4) is about 20% of the entire range of group averages (9.0 to 22.8) for ACT.] This may be due to the nature of the students who take the ACT, the states in which the ACT is used, and the populations of high schools in these states. It is certainly not the type of distribution which you would obtain from random samples of groups of 50 to 100 students.

Table 2: Pupil PR's for Stanford Large Group PR's, Stanford Small Group PR's and ITBS School Average PR's

Stanford Large Group		Large Group PR						
Grade	Test	99	90	75	50	25	10	1
2	Reading Comp	79	70	62	52	42	34	24
6	Reading Comp	76	66	58	48	37	31	21
8	Reading Comp	81	71	61	50	39	29	19
2	Spelling	79	68	60	50	40	33	22
6	Spelling	74	65	57.5	49.7	41	33	24
8	Spelling	71	64	58	51	44	38	30
2	Language (No Test)							
6	Language	77	66	57	47	38	28	18
8	Language	78	68	60	51	41	34	23
2	Language Total (No Total)							
6	Language Total	75	66	57	48	39	30	22
8	Language Total	74	61	52	41	30.5	23	14
2	Concepts of Number	82	70	62	51	40	31	19
6	Concepts of Number	77	65	55	45	34	26	16
8	Concepts of Number	80	69	61	50	41	32	21
2	Math Computation	80	69	61	51	41	33	21
6	Math Computation	79	67	59	46	34	27	16
8	Math Computation	80	69	61	51	41	33	21
2	Environment	82	70	61.5	52	42	33	21
6	Social Science	76	65	56	47	38	30	20
8	Social Science	83	73	64	52	39	30	19
2	See Environment Grade 2							
6	Science	80	67	56	47	37	28	18
8	Science	82	71	62	51	41	31	20

Stanford Small Group		Small Group PR						
Grade	Test	99	90	75	50	25	10	1
2	Reading Comp	88	76	66	53	40	30	16
6	Reading Comp	87	75	65	53	41	30	17
8	Reading Comp	88	74	62	48	33	23	10
2	Spelling	88	75	64	51	38	28	14
6	Spelling	86	74.5	63	52	40	30	17
8	Spelling	87	74	63	50	37	26	13
2	Language (No Test)							
6	Language	87	75	64	52	39	28	14
8	Language	84	71	60	47	34	24	12
2	Language Total (No Total)							
6	Language Total	88	76	67	54	41	30	16
8	Language Total	86	73	63	49	35	25	13
2	Concepts of Number	88	75	65	52	39	29	15
6	Concepts of Number	87	75	63	50	37	26	14
8	Concepts of Number	86	71	61	46	33	23	11
2	Math Computation	87	75	69	51	38	27	15
6	Math Computation	85	72	62	50	36	26	14
8	Math Computation	84	71	60	46	34	23	12
2	Environment	90	75	65	53	41	31	14
6	Social Science	80	76	66	54	41	30	16
8	Social Science	85	72	61	48	35	25	13
2	See Environment Grade 2							
6	Science	87	74	63	51	38	27	14
8	Science	86	73	61	48	35	25	13

Table 2: Pupil PR's for Stanford Large Group PR's, Stanford Small Group PR's and ITBS School Average PR's

ITBS Grade	Test	ITBS School Average PR's 99	90	75	50	25	10	1
2	Reading Comp	86	75	67	55	42	33	17
6	Reading Comp	88	74	64	53	41	30	16
8	Reading Comp	91	72	62	51	38	29	16
2	Spelling	84	73	65	55	43	33	14
6	Spelling	82	71	63	51	41	32	19
8	Spelling	84	71.5	61	51	42	34.5	23
2	Usage & Expression	84	71	62	51	39	28	16
6	Usage & Expression	88	75	70	52	40	30	18
8	Usage & Expression	90	76	65	51	40	30	16
2	Language Total	78	66	59	51	42	35	25
6	Language Total	88	75	65	53	41	32	19
8	Language Total	90	74	63	53	42	33	19
2	Math Concepts	80	65	58	49	38	31	17
6	Math Concepts	87	74	65	52	39	28	15
8	Math Concepts	85	72	61	49	37	25	15
2	Math Computation	85	70	60	49	37	28	14
6	Math Computation	87	72	60	50	38	26	12
8	Math Computation	89	71	61	50	37	27	13
2	Social Studies	86	75.5	64	52	40	30	17
6	Social Studies	82	71	61	49	35	24	14
8	Social Studies	80	70	59	49	35	23	12
2	Science	84	74	64	53	43	32	17
6	Science	81	72	61	50	39	30	17
8	Science	80	70.5	60.5	49.5	38.5	27	17

In the late 1940's, E. F. Lindquist discussed the differences among pupil and school norms and the need for such norms. A lengthy excerpt from his speech to a group of measurement experts is provided in Appendix A.

In the Lindquist discussion, the pupil and school samples were very large—at least 6,000 students per test and from 193 to over 400 schools. Lindquist obtained distributions of mean scores of random samples for these tests, which included just about everything for which Lindquist could obtain data, from the Iowa Every Pupil Tests of Basic Skills to the Cooperative subject achievement tests, then published by ETS. Lindquist concluding comments are provided below.

> In the first place, let me say that, in every case, the variability of school means is far larger than would be the variability of means of random samples of the same size drawn from the same total population. If, in any case, one were to make a simple analysis of the total variance into "between schools" and "within schools" components, one would find that the "between schools" variance is from two to ten times as large as that which could be accounted for by random sampling alone. The ratio would vary greatly, however, from test to test.
>
> In third-grade arithmetic (Iowa Tests of Basic Skills) the ratio of the 10th to the 90th percentile range for school averages and for pupil scores is exactly .5. For sixth-grade arithmetic it jumps to .55, but for third-grade language it is .36. For a high school physics test the ratio is .61, while for a United States history test at the same grade level it is .37. In general, the ratio seems to be largest where the opportunity to learn what is tested is relatively restricted to the classroom and to the formal school curriculum and to be least where the achievement measured might more often result from out-of-school experiences and informal learning.
>
> Well, I think I have said enough to make my point. School averages are less variable than pupil scores, but much more variable than the means of random samples. More important, the relative variability of school averages differs markedly from test to test and grade to grade. If school administrators, then, are to interpret properly measures of <u>school</u> achievement for purposes of curriculum evaluation or for the evaluation of group achievement in general, they must be supplied with norms specifically adapted to that purpose—i.e., they must have norms descriptive of distributions of school <u>averages</u> for defined populations of schools, just as for the interpretation of individual achievement they now have norms descriptive of distributions of pupil scores for defined populations of pupils."

Ratios similar to Lindquist's were extracted from Table 1 for the ITBS and Stanford. (Stanford Grade 2 "Social Studies" is called "Environment" and includes social studies and science.) All ratios of group to pupil ranges for these tests, including Stanford small group and large group ratios, are provided in Table 3.

TABLE 3

Ratios of 10th to 90th Percentile Ranges for Group Averages and Pupil Scores for the ITBS, the Stanford Achievement Test, and the American College Testing Program (ACT)

TEST	Spring Grade 2	Fall Grade 6	Spring Grade 8	Grade 11/12
ITBS Reading	.33	.48	.49	-
Stanford Reading (Small)	.49	.48	.53	-
Stanford Reading (Large)	.37	.36	.42	-
ITBS Spelling	.31	.43	.39	-
Stanford Spelling (Small)	.49	.46	.51	-
Stanford Spelling (Large)	.35	.31	.26	-
ITBS Language Total	.48	.48	.46	-
Stanford Language Total (Small)	-	.50	.50	-
Stanford Language Total (Large)	-	.36	.39	-
ITBS Math Computation	.43	.47	.47	-
Stanford Math Computation (Small)	.50	.48	.51	-
Stanford Math Computation (Large)	.37	.40	.38	-
ITBS Math Concepts	.37	.49	.52	-
Stanford Math Concepts (Small)	.49	.50	.52	-
Stanford Math Concepts (Large)	.40	.40	.38	-
ITBS Social Studies	.46	.49	.51	-
Stanford Social Studies (Small)	.51	.49	.50	-
Stanford Social Studies (Large)	.43	.36	.45	-
ITBS Composite	.51	.55	.57	-
Stanford Composite (Small)	.54	.44	.49	-
Stanford Composite (Large)	.45	.37	.37	-
ACT Composite	-	-	-	.32

Lindquist's observations were that tests that measured skills taught primarily in school would have the largest ratios. For ITBS, Language Total (spelling, punctuation, capitalization, and usage/expression) and Social Studies have ratios approaching .50 for all grades. ITBS Math Computation also has ratios approaching .50 except for the .43 ratio in second grade. Reading and Math Concepts follow a similar pattern but have ratios below .40 in the second grade. Spelling, which one might consider to be school related, has the lowest ratios.

The Stanford Spelling test also has the lowest ratios! The pattern of ratios on Stanford Math Computation is similar to the ITBS. However, the pattern for most other tests is not the same as the ITBS. The ratios for Reading and Social Studies drop slightly in Grade 6 on the Stanford, and the ratios for Stanford Math Concepts are about the same over all grades.

The ratios for the ITBS and Stanford Composite differ considerably and exhibit different patterns across grades. Exactly why this occurs is unknown. Very likely it is a function of the developmental score scales used and how these are derived.

Lindquist may have been dealing with raw score distributions. We don't know. None of the tests on the market have norms for pupils and groups which use raw scores. Pupil norms for tests. but not always totals. can be directly translated into raw scores because raw score to developmental score tables are included in the manuals. However. one can only estimate average raw score data for group norms because there is not a monotonic relationship between raw scores and developmental scores. A raw score "average" derived from a developmental score average using an RS-DS table for pupil scores is rarely the same as a raw score average obtained from the sum of raw scores divided by the N. Nevertheless, we did a raw score analysis for two tests, Reading and Spelling, for the ITBS and Stanford. The ratios for developmental scores (DS) and for raw scores (RS) are provided in the table below. The DS ratios come from Table 3.

Ratios of 10th to 90th Percentile Ranges
Group Averages and Pupil Scores for Developmental Scores (DS) and
Raw Scores (RS)

	Reading			Spelling		
	Gr. 2	Gr. 6	Gr. 8	Gr. 2	Gr. 6	Gr. 8
ITBS DS	.33	.48	.49	.31	.43	.39
ITBS RS	.48	.50	.54	.48	.41	.38
STANFORD (Small) DS	.49	.48	.53	.49	.46	.51
STANFORD (Small) RS	.60	.50	.58	.54	.49	.54
STANFORD (Large) DS	.37	.36	.42	.35	.31	.26
STANFORD (Large) RS	.37	.40	.46	.38	.36	.27

The ratios for ITBS raw scores are more consistent with Lindquist's observations and. in my opinion. more consistent with what one would expect. The pattern for the Stanford is quite similar to the ITBS pattern except for the Grade 2 Reading small group ratio of .60. This ratio is probably inflated because of the non-monotonic relationship between raw scores and developmental scores noted in the paragraph which precedes the table.

EFFECTS OF EXCLUDING STUDENTS FROM AVERAGES

One of the more interesting aspects of the state data reported in the Friends For Education (FFE) report was the total student enrollment for which state averages were provided. The FFE report provided N's tested, only. Although final state enrollment figures for the 1986-1987 school year are not yet available from the National Center of Educational Statistics, projections of enrollments based on 1985-1986 statistics indicate that, in relation to total enrollment, the percentage of students for whom averages were reported ranged from about 86% to 97% for most states. In one state, the percent was about 80! The nature of the population of students not tested, or at least not reported, could have a dramatic effect on the scores reported.

In order to investigate the effect of systematic and non-systematic exclusion of students from reports, we obtained scores from norms tables for the 95th, 90th, 10th, and 5th percentiles. We then calculated the effect of excluding either 5% or 10% of the students at these score levels to determine the effect of systematic exclusion of students. For example, if schools were to say, "We don't want to score too high," they might exclude 5% or 10% of their students centered around the 90th or 95th percentiles. (Essentially, 10% of the students around the 95 percentile is all of the top 10% of the students.) Unlikely as it may seem, other schools might say, "We don't want to score too low." Thus, these schools might exclude many low-scoring students from their reports.

Regardless of how great an effort is made to do make-up testing, especially in state programs, some students are not included in the state averages. Since we had to obtain extreme scores for the systematic exclusion analysis, we used 2.5% of the students at each extreme (95th percentile and 5th percentile or 90th percentile and 10th percentile) to estimate the effect on averages for non-systematic exclusion of pupils. The procedure we used does not really measure non-systematic exclusion of students because we systematically excluded students only at the extreme percentiles.

For the ITBS GE and CAT-E scaled score, we obtained mean scores and scores at the 95th, 90th, 50th, 10th, and 5th percentiles from published tables. For Stanford, mean scale scores were not available at the time this analysis was done (February 28), but scale scores were obtained for the five percentiles. All tabled pupil PR's were read to the nearest tenth. Mean scores were obtained from ITBS and CAT technical manuals. We then calculated the new mean using the formula:

Mean score - (% Excluded)(Score at Exclusion Level) ÷ (1.00 - Decimal value of %)

For ITBS Grade 2 Reading Comprehension for 10% excluded at the 95th percentile, the calculation would be:

28.8 - (.10)(49.0) ÷ (1.00 - .10) = 26.56

The pupil PR for the GE of 26.56 was then obtained from existing norms tables to obtain the pupil PR if 10% of the students around the 95th percentile were excluded.

The table below shows the data for all calculations for ITBS, Grade 2 Reading Comprehension, Spring. Data for the national standardization average are on the far right.

% Excluded	@ PR	Score for this PR	New Average	Pupil PR of New Avg.	Nat'l Avg. 28.8
10%	95	49.0	26.56	45	
5%	95	49.0	27.74	50	Pupil PR of
10%	90	44.0	27.11	47	Nat'l Avg.
5%	90	44.0	28.00	51	53
	50	27.8	--	--	
10%	10	15.0	30.33	58	
5%	10	15.0	29.53	55	
10%	5	12.0	30.67	59	
5%	5	12.0	29.68	56	
5%	90-10	See	28.72	53	
5%	95-5	Above	28.61	53	

If a school mirrors the national distribution, the mean would be 28.8 and the median would be 27.8. For Stanford, at the time the calculations were done, we did not have the scale score means so we had to use medians. (This portion of the paper was completed on February 27. The Stanford means shown in Table 4 were received March 7, but we didn't recalculate using means.) However, using standard procedures for estimating means from existing distributions, we estimated that Stanford means and medians are very nearly the same, probably more similar than the ITBS means and medians. In order to estimate how large a difference one would obtain if medians were used instead of means, we also calculated all ITBS data using medians instead of means. The average difference between the two sets of data was less than a half of a PR, and in no case was it greater than one PR. The PR's were obviously different, but amount of change in PR was the same, regardless if means or medians were used. Hence, we are reasonably certain the change in PR's shown for Stanford data is comparable to what would have been obtained had we had Stanford means.

Table 4 lists the scores used in the calculations. It also lists the median scores for ITBS and CAT. The ITBS median scores could be used to check the statements made in the preceding paragraph. The table also shows the differences between mean and median scores for ITBS and CAT. Some of the CAT differences appear to be large, but the total range of scores needs to be considered when these differences are compared to other tests, such as the ITBS. For example, the score range from the 5th to 95th percentile for ITBS Reading Comprehension, Grade 2 Spring, is 12 to 49 or 35 points. The difference between the mean and median is 1 point. For the comparable CAT test the score range from the 5th to the 95th percentile is 494 to 729 or 235 points and the difference between the mean and median is 11 points, which is more or less comparable to the ITBS.

Table 5 lists the changes in percentile ranks for systematic and non-systematic exclusion of students. There were really no effects for non-systematic exclusion, even with the extreme scores we used in our calculations. The five differences of +2 all were rounded upward from actual values of +1.5 to +1.8.

Systematic exclusion of students yields very different changes in PR's depending on the test battery, test, grade, and the score level of excluded students. The comparable pairs of changes are designated with an a, b, c, or d at the top of each column. For example, 10% excluded at PR of 95 and PR of 5 are comparable pair a. The 10% exclusion always effects the greatest change in PR's, as one would expect, but the size of the differences varies by test battery for comparable pairs. For the ITBS, the change is consistently greater when you exclude better students than when you exclude comparable low-scoring students. For CAT, the change is greatest when you exclude low-scoring students. For Stanford, the differences vary somewhat. In Grade 2 and Grade 6 Stanford, the change is slightly greater when you exclude high-scoring students. In Grade 8 the changes are virtually identical for all comparable pairs.

The magnitude of the changes that occur when you systematically exclude students also varies by test battery. On ITBS, the maximum changes are -11 and +10 PR's. Ninety-five percent of the changes are between -7 to -2 and +2 to +7 PR's. On CAT the maximum changes are -10 and +13 PR's, and ninety-five percent of the differences are between -7 to -2 and +2 to +9. On Stanford maximum changes are -8 and +8 PR's and ninety-five percent of the changes are between -7 and -3 and +3 to +7.

All of these changes are in terms of pupil percentile ranks. It is obvious that, if group norms were used, the changes in group percentile ranks would be significantly greater for most tests in most grades.

Table 4

Grade Equivalent Scores Which Correspond to Selected PR and the Mean Grade Equivalent Score – ITBS Forms G–H

Grade 2 Spring	PR=95	PR=90	PR=50	Mean	PR=10	PR=5
Reading Comp	49	44	27.8	28.8	15	12
Spelling	50	45	28	29.4	16	13
Usage & Expression	57	51.5	29.5	30.7	12	K.9
Math Concepts	44	40.5	29	29.5	18.7	16.5
Math Computation	42	39	29	29.1	19.5	17.0
Science	60.5	54	28	29.6	K8.5	K5
Social Studies	62.5	56	28	31	11	K7
Complete Composite	44.5	41	28.5	29	17.7	15.5

Grade 6 Fall	PR=95	PR=90	PR=50	Mean	PR=10	PR=5
Reading Comp	90.5	84	62	62.2	40.5	36
Spelling	92	86	62	62.1	39	32
Usage & Expression	94.5	88	62	61.8	35	30
Math Concepts	86.5	81	61.7	62.6	45	40.5
Math Computation	79.5	74.5	61.7	61.5	47	41.5
Science	104	97.5	62	63.1	32	25.5
Social Studies	101.5	94	62	62.5	33	27
Complete Composite	85	80	61.3	61.8	44	40.5

Grade 8 Spring	PR=95	PR=90	PR=50	Mean	PR=10	PR=5
Reading Comp	119.5	112	88	86.6	59	51.5
Spelling	120	114	87	85.1	53	43.5
Usage & Expression	122	115	88	86	54	45.5
Math Concepts	117	112	89	88.4	64	58
Math Computation	106.5	101.5	87	85.1	65	56.5
Science	131	125	87	86.6	48	40.5
Social Studies	128	120	87	85.4	50	40.5
Complete Composite	113.5	108	87	86.3	63	58

Table 4

Scaled Scores Which Correspond to Selected PR and the Mean Scaled Score – CAT, Forms E – F

Grade 2 Spring	PR=95	PR=90	PR=50	Mean	PR=10	PR=5
Reading Comp	729	709	640	629	534.5	494
Spelling	704	688	642	638	577.5	554
Language Expression	746	742.5	667	656	550	512
Math Concepts & App.	761.5	719	643	647	576.5	551.5
Math Computation	734.7	732.5	635	636	539.5	505
Science	708.5	679.5	591	590	490	460
Social Studies	731.5	696	615	616	534.5	499.5
Total Battery	710	699.5	647	640	565.5	535

Grade 6 Fall	PR=95	PR=90	PR=50	Mean	PR=10	PR=5
Reading Comp	773.5	764	730.5	727	684	666
Spelling	770.5	759.5	724.5	722	681	664
Language Expression	778	761	708	703	633	597.5
Math Concepts & App.	767	755.5	714.5	710	657.5	635
Math Computation	787.5	779	745.5	742	698	676.5
Science	759	739	675	NA*	574	523
Social Studies	746.5	734.5	691	NA*	625.5	587
Total Battery	769	759.5	721	718	668.5	650

Grade 8 Spring	PR=95	PR=90	PR=50	Mean	PR=10	PR=5
Reading Comp	789.5	780	757.3	755	725	713
Spelling	789.5	779	749.3	748	715.5	700.5
Language Expression	817	797	737	734	664.5	624.5
Math Concepts & App.	813	800.5	755.5	754	705	683.5
Math Computation	845.5	833	797.5	792	739.5	718.5
Science	785	771.5	724	715	638	598.5
Social Studies	790.5	776	724	715	630.5	595
Total Battery	806	796	756	754	707.5	686.5

*NA – Not available at the time this paper was prepared.

Table 4

Scaled Scores Which Correspond to Selected PR – Stanford 7

Grade 2 Spring	PR=95	PR=90	PR=50	Mean	PR=10	PR=5
Reading Comp	657	640	577	577.4	515	497
Spelling	653	635	572.5	572.3	509	491
Concepts of Number	636.5	620.5	564	564.2	507	490.5
Math Computation	651.5	634	572.5	572.9	511	494
Environment	641	630	587	586.7	544	533
Complete Battery	628	615.5	573	573.5	530.5	519

Grade 6 Fall	PR=95	PR=90	PR=50	Mean	PR=10	PR=5
Reading Comp	732	715	654.5	655.4	593.5	577
Spelling	749	730	664.5	664.9	599	581
Language	721.5	707.5	659	659.8	611.5	598.5
Concepts of Number	712.5	701.5	651	651.4	600.5	586.5
Math Computation	720.5	706	652	652.8	599	594.5
Science	709	695.5	648	648.6	600.5	588
Social Science	705.5	691.5	641	641.2	590.5	576.5
Complete Battery	702	690.5	650	650.8	610.5	599

Grade 8 Spring	PR=95	PR=90	PR=50	Mean	PR=10	PR=5
Reading Comp	756.5	739	679.5	679.5	619	601.5
Spelling	782	764	702.5	702.9	641	623
Language	755.5	740.5	686	686.6	632.5	616.5
Concepts of Number	786	768	705.5	705.9	643	625
Math Computation	785.5	768	706.5	706.6	643	625
Science	742.5	727.5	677	677.2	626.5	610.5
Social Science	755.5	739	682	682.8	626	609.5
Complete Battery	746	733.5	686	685.8	639.5	625.5

The Stanford means shown in this table were received too late to do the calculations for Table 5.
(See page 16)

Table 5

Percentile Rank Changes When 5%, or 10% of Students Are Excluded at the Indicated Percentile Ranks and 5% Are Excluded From Both Ends of the Distribution, ITBS, Forms G – H

	(a) 10% @ PR	(b) 5% @ PR	(c) 10% @ PR	(d) 5% @ PR	(c) 10% @ PR	(d) 5% @ PR	(a) 10% @ PR	(b) 5% @ PR	5% Top and
Grade 2 Spring	95	95	90	90	10	10	5	5	Bottom
Reading Comp	−8	−3	−6	−2	+5	+2	+6	+3	0
Spelling	−8	−4	−6	−3	+4	+2	+5	+3	0
Usage & Expression	−6	−3	−5	−2	+5	+2	+6	+3	+2
Math Concepts	−7	−4	−6	−2	+5	+2	+7	+3	0
Math Computation	−8	−4	−6	−3	+6	+2	+7	+3	−1
Science	−8	−4	−7	−3	+6	+3	+6	+4	0
Social Studies	−10	−4	−8	−3	+4	+2	+5	+2	0
Avg for 7 Tests	−8	−4	−6	−3	+5	+2	+6	+3	+.1
Complete Composite	−8	−3	−6	−3	+5	+2	+6	+3	0
Grade 6 Fall									
Reading Comp	−6	−3	−5	−2	+6	+3	+7	+3	0
Spelling	−7	−3	−6	−2	+6	+3	+7	+4	0
Usage & Expression	−6	−3	−5	−3	+4	+2	+5	+3	0
Math Concepts	−11	−3	−6	−2	+5	+2	+7	+3	0
Math Computation	−7	−3	−5	−2	+5	+3	+7	+3	0
Science	−7	−3	−6	−3	+4	+1	+5	+2	0
Social Studies	−8	−4	−6	−4	+5	+3	+6	+3	0
Avg for 7 Tests	−7	−3	−6	−3	+5	+2	+6	+3	0
Complete Composite	−7	−4	−6	−3	+5	+2	+6	+3	0
Grade 8 Spring									
Reading Comp	−6	−5	−7	−4	+5	+1	+6	+2	0
Spelling	−6	−3	−5	−2	+5	+2	+7	+3	0
Usage & Expression	−5	−2	−4	−2	+4	+2	+6	+3	0
Math Concepts	−7	−4	−6	−3	+6	+2	+7	+3	0
Math Computation	−5	−3	−4	−2	+7	+3	+10	+5	+1
Science	−7	−3	−6	−3	+4	+1	+5	+2	0
Social Studies	−8	−4	−6	−4	+5	+3	+6	+3	+.1
Avg for 7 Tests	−6	−3	−5	−3	+5	+2	+7	+3	0
Complete Composite	−8	−4	−6	−4	+5	+3	+6	+3	0

The a, b, c, and d designations above the columns refer to comparable pairs of scores at the top and bottom of the distribution. For example, (a) is for 10% of students excluded around the 95th and 5th percentiles.

Table 5 (Continued)

Percentile Rank Changes When 5%, or 10% of Students Are Excluded at the Indicated Percentile Ranks and 5% Are Excluded From Both Ends of the Distribution, CAT, Forms E – F

	(a) 10% @ PR	(b) 5% @ PR	(c) 10% @ PR	(d) 5% @ PR	(c) 10% @ PR	(d) 5% @ PR	(a) 10% @ PR	(b) 5% @ PR	5% Top and
Grade 2 Spring	95	95	90	90	10	10	5	5	Bottom
Reading Comp	−7	−4	−6	−3	+7	+3	+11	+4	0
Spelling	−6	−3	−5	−3	+7	+3	+9	+4	+1
Language Express	−6	−3	−6	−2	+8	+4	+11	+5	+1
Math Conc & App	−10	−4	−7	−3	+6	+3	+8	+3	0
Math Computation	−8	−1	−8	−1	+5	+3	+6	+4	0
Science	−8	−3	−6	−3	+8	+4	+9	+5	+1
Social Studies	−10	−5	−8	−3	+6	+3	+8	+4	0
Avg for 7 Tests	−8	−3	−7	−3	+7	+3	+9	+4	+.4
Total Battery	−6	−3	−5	−3	+6	+3	+9	+4	+1
Grade 6 Fall									
Reading Comp	−6	−3	−5	−3	+7	+3	+10	+4	0
Spelling	−5	−2	−4	−1	+7	+3	+9	+5	0
Language Express	−7	−4	−5	−3	+7	+3	+10	+5	0
Math Conc & App	−6	−3	−5	−2	+7	+3	+9	+4	0
Math Computation	−7	−3	−6	−3	+6	+2	+9	+4	0
Science*	−6	−3	−4	−2	+8	+4	+13	+6	+2
Social Studies*	−7	−2	−5	−1	+7	+4	+12	+6	+2
Avg for 7 Tests	−6	−3	−5	−2	+7	+3	+10	+6	+.5
Total Battery	−6	−3	−5	−2	+7	+4	+10	+5	+1
Grade 8 Spring									
Reading Comp	−8	−4	−7	−3	+6	+3	+9	+3	0
Spelling	−9	−3	−6	−3	+7	+3	+8	+5	0
Language Express	−7	−3	−5	−2	+7	+4	+10	+5	0
Math Conc & App	−8	−4	−3	−3	+5	+3	+8	+4	0
Math Computation	−6	−3	−5	−2	+8	+4	+11	+5	+2
Science	−7	−4	−5	−3	+7	+3	+11	+5	+1
Social Studies	−6	−3	−5	−2	+8	+4	+12	+5	+2
Avg for 7 Tests	−7	−3	−5	−3	+7	+3	+10	+5	+.7
Total Battery	−7	−4	−6	−3	+5	+3	+8	+4	0

*These calculations are based on median scores, not mean scores, because mean scores were not available at the time of publication of this paper.

Table 5 (Continued)

Percentile Rank Changes When 5%, or 10% of Students Are Excluded at the Indicated Percentile Ranks and 5% Are Excluded From Both Ends of the Distribution, Stanford 7

Grade 2 Spring	(a) 10% @ PR 95	(b) 5% @ PR 95	(c) 10% @ PR 90	(d) 5% @ PR 90	(c) 10% @ PR 10	(d) 5% @ PR 10	(a) 10% @ PR 5	(b) 5% @ PR 5	5% Top and Bottom
Reading	−7	−3	−6	−2	+6	+2	+7	+3	0
Spelling	−7	−4	−5	−3	+6	+3	+8	+4	0
Math Concepts	−7	−4	−5	−3	+5	+2	+7	+3	0
Math Computation	−7	−4	−5	−3	+5	+3	+7	+4	0
Environment	−6	−3	−5	−2	+5	+2	+6	+3	0
Avg. for 5 Tests	−7	−4	−5	−3	+5	+2	+7	+3	0
Complete Battery	−8	−4	−6	−3	+6	+2	+7	+3	0
Grade 6 Fall									
Reading	−7	−4	−5	−3	+6	+3	+7	+4	0
Spelling	−8	−4	−6	−3	+6	+3	+7	+4	0
Language	−8	−4	−6	−4	+5	+3	+7	+3	0
Math Concepts	−7	−3	−6	−3	+6	+3	+7	+3	0
Math Computation	−8	−4	−6	−3	+6	+3	+6	+3	0
Science	−7	−3	−5	−2	+5	+3	+7	+3	0
Social Science	−7	−3	−6	−3	+4	+3	+7	+3	0
Avg. for 7 Tests	−7	−4	−6	−3	+5	+3	+7	+3	0
Complete Battery	−8	−4	−6	−3	+5	+2	+7	+3	0
Grade 8 Spring									
Reading	−7	−4	−6	−3	+6	+3	+7	+4	0
Spelling	−7	−4	−5	−3	+5	+3	+7	+4	0
Language	−8	−4	−6	−3	+6	+3	+8	+4	0
Math Concepts	−7	−4	−5	−3	+5	+3	+7	+4	0
Math Computation	−7	−4	−5	−3	+6	+3	+8	+4	0
Science	−7	−3	−6	−3	+6	+3	+7	+4	0
Social Science	−8	−4	−6	−3	+5	+3	+7	+4	0
Avg. for 7 Tests	−7	−4	−6	−3	+6	+3	+7	+4	0
Complete Battery	−7	−3	−5	−2	+5	+2	+7	+3	0

Standford calculations are based on median scores, not mean scores because mean scores were not available at the time of publication of this paper.

RELATIONSHIPS AMONG DIFFERENT KINDS OF SCORES AND DEMOGRAPHIC DATA

The FFE report cited a number of purported correlates of test scores in arguing that certain states' test scores were grossly inflated. In Illinois, we have a report card that lists a number of school characteristics, including low-income percent, graduation rate, and other data including test scores. One newspaper article was headlined, "Poverty is no excuse for ACT failure."

Most educators realize that these various correlates of achievement don't correlate too well with test scores, especially when students not included in the aggregations differ, poverty indices vary by state, etc.

Newspapers throughout the State of Illinois reported on the report card. I originally intended to analyze data by school, but "school" data reported for some characteristics were really district data. We ended up with the following data for each of 85 Chicago area districts, 52 of which were one building districts:

- Number tested, ACT Composite average, building PR and NCE of this PR, and pupil PR and NCE of this PR.

- For graduation data, low income percent, and percent of class tested we had %'s, group PR, NCE of group PR, and rank among the 85 districts.

I hypothesized that building PR would correlate with various school characteristics better than the pupil PR because of the range restriction of pupil PR's.

Local distributions of the graduation rate, low income percent of students, and percent of students tested were developed. The "local PR" was then calculated and this PR was converted to an NCE. For graduation rate, the lowest PR and, hence, the lowest NCE is for a district with the lowest graduation rate—i.e., the greatest dropout rate. For low income percent, the lowest PR and lowest NCE is for the district with the highest percent of low income students. For percent of students who took the ACT, the lowest PR and lowest NCE is for the district in which the lowest percent of students were tested. The ranks are explained in Table 6.

Note: We began with a total of 85 districts. All 85 districts were used to develop the local distributions of data from which the NCE's for graduation percent, low income percent, and class tested percent were obtained. The last district on the file was inadvertently dropped from final analysis. Fortunately, this district was very nearly average for the group—the ACT average was 19.4 and NCE's were very nearly 50 on the other characteristics. This accounts for the NCE means that are slightly greater than 50. The error was not noticed until the tables were built and most comments were written—late 3/28. Sorry.

The mean score for the 32,695 students in 84 districts analyzed was estimated by summing the products, N times average score of the district, and dividing by 32,695. The ACT average for this group was 20.09, which non-unexpectedly is at about the 56th percentile on ACT pupil norms and 1.4 points greater than the national ACT average.

Table 6 summarizes the data for the 84 districts. We statisticized PR's because we wanted to see how well these non-equal-interval scores in a fairly normal distribution would compare with other scores. 'Tain't too professional by test standards standards, but a lot of actions aren't purely professional.

Table 6
Variables Used and Summary Statistics for These Variables

Variable	Median	Mean	SD	Range
NCNT = # of Students per District	246.0	389.22	406.00	16-2339
AVG = ACT Composite Score Average	19.55	19.54	1.67	13.9-22.6
BLDG PR = Building PR of ACT Average	57.0	57.50	27.61	4-98
BLDG NCE = Building NCE of ACT Average	53.5	55.33	18.75	13-93
PUPIL PR = Pupil PR of ACT Average	54.0	53.73	9.52	23-72
PUPIL NCE = Pupil NCE of ACT Average	52.0	52.00	5.22	34-62
GRAD PCT = % of Students Who Graduate	89.4	87.16	29.25	50.0-99.1
GRAD NCE = NCE of This Group *	50.0	50.04	21.07	1-99
GRAD RANK = Highest Grad. % is Ranked 1	--	--	--	1-84
INC PCT = % of Low Income Students	4.55	7.53	8.05	0.0-33.9
INC NCE = NCE of This Group *	48.5	50.12	20.87	1-99
INC RANK = Highest % of INC PCT is Ranked 1	--	--	--	1-84
CLS PCT = % of Class Tested	60.35	60.63	14.42	32.7-98.1
CLS NCE = NCE of This Group *	50.5	50.21	20.93	1-99
CLS RANK = Lowest % Tested is Ranked 1	--	--	--	1-84

* See Narrative

The national building PR's for these 84 districts range from 4 to 98. The pupil PR's range from 23 to 72. The number of students tested in the 84 Chicago "collar" districts varies from 16 to 2339. The means and medians for scores, graduation rate, and class percent tested are quite similar. The percent of low income means and medians differ considerably. The kurtosis and skewedness for these three characteristics—based on all 85 districts—were: Graduation, k=3.12, s=-1.46; Income, k=2.25, s=1.70; and % Tested, k=-.50, s=.30.

Correlations among scores and characteristics are reported in Table 7. (Since we forgot to wipe out NCNT, we decided to report the all-important correlation between NCNT and scores/characteristics, which were significant at the .01 level with everything except low income %! We also included Graduation percentile rank (PR), Income PR, and Percent Tested PR in the analysis. These aren't shown in Table 7, but they were of the same magnitude as the correlations with the respective NCE's in all instances!)

General observations about Table 7 are:

1. Building PR or NCE does not have a higher correlation than pupil PR or NCE with other variables, but average score might just as well be used.
2. NCE's or similar transformations of local ranks for non-normally distributed characteristics correlate somewhat better with test performance than the raw statistics, but a larger random sample is needed to ascertain if this is really so.
3. Although I know of some instances where students are advised not to take the ACT, perhaps so as to not lower the school average on the "report card", the correlation between N tested and test scores doesn't seem to indicate the practice is widespread or at least differs by school.

Table 7--Correlations Between ACT Scores and Various Characteristics

	Avg.	Bldg. PR	Bldg. NCE	Pupil PR	Pupil NCE	Grad. Data %	Grad. Data NCE	Grad. Data Rnk	Low Income Data %	Low Income Data NCE	Low Income Data Rnk	Class % Tested %	Class % Tested NCE	Class % Tested Rnk
NCNT	.322	.335	.337	.312	.306	.276	.287	-.270	-.188*	.307	-.333	.422	.418	.434
AVG		.978	.987	.999	.998	.616	.604	-.577	-.618	.671	-.645	.692	.691	.680
BLDG. PR			.989	.978	.970	.585	.583	-.568	-.585	.658	-.645	.707	.699	.700
BLDG. NCE				.988	.984	.600	.610	-.583	-.583	.671	-.647	.719	.712	.704
PUPIL PR					.998	.615	.604	-.576	-.613	.670	-.643	.688	.687	.675
PUPIL NCE						.620	.611	-.583	-.622	.673	-.644	.686	.687	.673
GRAD PCT							.943	-.919	-.530	.540	-.515	.648	.672	.655
GRAD NCE								-.984	-.496	.577	-.534	.690	.701	.689
GRAD RANK									.484	-.544	.509	-.677	-.683	-.682
INC PCT										-.882	.855	-.520	-.538	-.512
INC NCE											-.983	.660	.662	.640
INC RANK												-.645	-.646	-.629
CLS PCT													.992	.985
CLS NCE														.984

* Not significant at the .01 level.

IMPLICATIONS FOR INTERPRETATION AND REPORTING

It should be clear from Table 3 that the pupil median does not correspond to the median of group averages very often. The mean of pupil scores from the national standardization sample typically does not correspond to either of these. For example, for ITBS Grade 2 Spring Reading Comprehension, the median GE pupil score is 27.8, the mean score of all pupils in the national standardization sample is 28.8, and the median school score is 29.4. For Grade 6 Fall the numbers are 62, 62.2, and 63.4, respectively.

After the Friends for Education report was published it became quite clear to me that the relationship among pupil and group scores is such that, when single point indicators are reported, group norms ought to be used. Unfortunately, most publishers don't have such norms. With the advent of annual norms--the latest Madison Avenue promotion in the industry--data for schools and perhaps even school districts could be obtained. Annual norms are based on user data, weighted to be representative of the nation as a whole, but include all students (more or less) in all buildings and the district.

The "more or less" in parentheses is an important factor in the goodness, badness, and representativeness of user norms. It is quite clear from the data presented that systematic exclusion of students can result in some very significant changes in pupil percentile ranks and these vary considerably across tests and grades.

The FFE report was titled, "Nationally Normed Elementary Achievement Tests in America's Public Schools: How All Fifty States are Above the National Average." The report focused on what was being reported. Most states are reporting that they are above the national pupil median, not necessarily the national average. Averages and medians are not the same. There are no state norms, but I have estimated that about 80% of the states probably have averages that are within 5 pupil PR's of the national average--not necessarily the pupil median. (The FFE report did not really indicate all fifty states were above average. For the states for which data are reported, the states were above the national median.) Systematic exclusion of students could easily move a state from -2 to +5 or +2 to -5 relative to the national median or mean. For example, for the ITBS and Stanford, for which median scores for groups are quite consistently greater than the pupil median, exclusion of low scoring students from group data could easily result in changing a pupil PR of 50 to 55. However, a pupil PR of 55 might still be a group PR of 50 or less!

Last, the so-called correlates of achievement are imperfect correlates. We did some scatterplots that showed a sufficient number of outliers to question how consistent and accurate graduation (or dropout) statistics and even low income statistics are, even within a state. Proportion of students tested relative to other characteristics is suspect in some schools. For example, one of the lowest proportion of students tested was in a district (building) with a 2.9% low income group, well above average, and a graduation rate of 73%—well below average, provided everybody's statistics were compiled the same way. The ACT score was 19.2, above the national average. The low income percent of students for all other schools testing fewer than 45% of the students in the class was greater than 10%!

CONCLUDING COMMENTS

These concluding comments are essentially the same as the comments I've made many times. (1) Single-point indicators of achievement have little, if any, value to instructional planners. (2) Since we are willing to honestly report to pupils and parents that a pupil is at the 1st or 10th percentile, we should honestly report group data—and this can only be done by using group norms, not pupil norms. (3) Single-point indicators are often misinterpreted and misused, particularly by reporters. (4) "Correlates" of achievement are rather imperfect but probably vary more from state to state than within state.

If school personnel use pupil norms with average data, it seems advisable to adjust the pupil PR an amount equal to the difference between the scores that correspond to the pupil mean and the pupil median. If the mean-median difference is a +5 PR's, all pupil PR's ought to be adjusted down by .13 z-scores. Thus a mean developmental score that yields a pupil PR of 55 would be adjusted downward to a pupil PR of 50, a pupil PR of 75 would become a PR of 71, and a pupil PR of 90—which is about as high as you'll find—would be adjusted down to pupil PR of 88.

Publishers of standardized tests provide a considerable number of score conversion tables and large body of technical data for their tests. The essential test standards are met or exceeded for most tests published by most commercial publishers. This is not the case for many locally developed or state tests, even when such tests are used for retention in grade, the withholding of diplomas, or sanctions against schools! Such lack of adherence to test standards set forth by APA-AERA-NCME are unconscionable.

Why the double standard? I don't know. Perhaps bureaucracies can do what they wish, good testing practices not withstanding. If states use standard scores for reporting, the raw score-standard score-pupil PR tables ought to be published. If states use school average scores to determine eligibility for sanctions, norms for school averages should be published. The scores and their derivation should be described. Cautions regarding the interpretation and use of the various scores, averages, and other data ought to be published. Last, relationships among various data (correlations, at least) and "minor" statistics such as reliabilities (even KR-21's, if KR-20's are overwhelming) ought to be provided.

ACKNOWLEDGEMENT

The author wishes to acknowledge the considerable assistance of Phyllis Hammond and Carol Zillmann for extracting and organizing information, to Debra Timpe and Russell Barcelona for cranking out the data, and his wife, Carol, for allowing him to work on the paper during a "get-away-from-it-all" vacation.

Appendix A

Speech by E. F. Lindquist, Reproduced from Testing Problems in Perspective, A. Anastasi, American Council on Education, 1966, pp. 269-271.

the American Council on Education book entitled, Testing Problems in Perspective, pages 269-271, which was edited by Anne Anastasi.

"What I have to say is not at all profound, and I should be able to say it in just two or three mintues. As elementary as the observation I have to make, however, judging from current test standardization practices and from the literature on testing, it is one that is sorely in need of stressing even in as sophisticated a group as this one.

What I wish to do is to put in a plug for a special type of norm for educational achievement tests. It is a type of norm which, in consideration of the major uses generally made of achievement tests, should always be provided with tests of that character, but one which in actuality is very rarely provided.

Let me remind you first of two major uses which we generally claim for educational achievement tests. First, we claim that these tests will provide the teacher and counselor with a more adequate and dependable basis for the educational guidance of the individual pupil and for the individualization of instruction. Second, we claim that these tests will be useful to the school administrator as a partial basis for evaluating the content and methods of instruction, or in curriculum evaluation in general.

Now with reference to the first of these purposes, quite obviously we need norms that are descriptive of the distributions of measures of achievement of pupils. We interpret a pupil's score by comparing it with the scores of other pupils in the relatively homogeneous population to which he belongs. The second purpose, however, is the evaluation of group, not of individual, achievement. What the principal wants to know is, "How are we doing as a school?" or "How does this grade group compare in average achievement in relation to other similar grade groups?" Quite obviously, what is needed here is a type of norm that is descriptive of distributions of measures of group achievement, i.e., of distributions of school averages, not of distributions of individual pupil scores.

Since such norms are rarely provided, what the school principal does is to attempt to interpret his grade averages or grade medians in terms of the norms that are provided--norms descriptive only of distributions of pupil scores. What he really does, without often knowing that he is doing so, is to assume that the school averages are distributed just as are individual pupil scores. School averages, of course, are much less variable than pupil scores (although not as much less as many may assume), and he immediately makes wrong interpretations of his results on that account. A school average that coincides with the 75th percentile in the pupil score distribution, for example, might be a remarkably high average, but since it is "at the 75th percentile," it might not be regarded as high by the principal. What he does would not be so serious, however, if there were a constant relationship between the variability of school averages and the variability of pupil scores. In that case he could at least tell on what tests his school did well in relation to other schools, and could rank the tests in order of the relative performance of his school, even though he would be misled as to the true relative performance on any one test. Actually, however, the relationship is not constant, but varies widely from subject to subject and from one grade level to another.

I have with me here some data concerning the relative variability of individual pupil scores and of school means for 34 widely used achievement tests. I am not going to read all of these data to you but will just give you a sample of them. They are all based on large numbers of cases--in no instance upon less than 193 schools, in some cases as many as 400, and in no case less than 6,000 pupils--so the results are quite stable so far as sampling errors are concerned.

The tests include tests of elementary school achievement, such as the Iowa Tests of Basic Skills, and tests of high school achievement, such as the Cooperative subject achievement tests.

In the first place, let me say that, in every case, the variability of school means is far larger than would be the variability of means of random samples of the same size drawn from the same total population. If, in any case, one were to make a simple analysis of the total variance into "between schools" and "within schools" components, one would find that the "between schools" variance is from two to ten times as large as that which could be accounted for by random sampling alone. The ratio would vary greatly, however, from test to test.

In third-grade arithmetic (Iowa Tests of Basic Skills) the ratio of the 10th to the 90th percentile range for school averages and for pupil scores is exactly .5. For sixth-grade arithmetic it jumps to .55, but for third-grade language it is .36. For a high school physics test the ratio is .61, while for a United States history test at the same grade level it is .37. In general, the ratio seems to be largest where the opportunity to learn what is tested is relatively restricted to the classroom and to the formal school curriculum and to be least where the achievement measured might more often result from out-of-school experiences and informal learning.

Well, I think I have said enough to make my point. School averages are less variable than pupil scores, but much more variable than the means of random samples. More important, the relative variability of school averages differs markedly from test to test and grade to grade. If school administrators, then, are to interpret properly measures of <u>school</u> achievement for purposes of curriculum evaluation or for the evaluation of group achievement in general, they must by supplied with norms specifically adapted to that purpose--i.e., they must have norms descriptive of distributions of school <u>averages</u> for defined populations of schools, just as for the interpretation of individual achievement they now have norms descriptive of distributions of pupil scores for defined populations of pupils."

How are such norms developed? Quite simply, by obtaining frequency distributions of school average data from the national standardization program, plotting and smoothing these distributions, and reading the percentile ranks which correspond to various averages. I suspect that some school average norms are obtained from user data after a standardization using a pre-prepared model which simulates the national "school population." I have no quarrel with such a procedure provided, of course, that the model is developed before, not after, the school data are in-house.